OBELICS TASK 3.4.1 D-ANA
Data Analysis and Interpretation
Task Summary

Bojan Nikolic
University of Cambridge

2nd ASTERICS-OBELICS Workshop
16-19 October 2017, Barcelona, Spain.

H2020-Astronomy ESFRI and Research Infrastructure Cluster
(Grant Agreement number: 653477).

Task Scope

3.4.1 Data analysis and mining on PetaByte-scale datasets

- Statistically robust approaches for cross-matching between different instruments
- Multi-dimensional / Multi-resolution image joint analysis from different instruments
- Likelihood reconstruction with maximum efficiency on new computing technology

3.4.2 Workflow architectures for orchestration, A&A -- More from Sonia Zorba shortly

Participants

INAF	U. Cambridge	LAPP	IAP
APC	CPPM	CEA	ASTRON
	JIVE	INFN	

Overlap of people and s/w projects between the OBELICS tasks

Diverse Use Cases -- Common Themes

Scientific

- Statistical efficiency
  - Energy efficiency
  - Data efficiency
  - Follow-up efficiency
- Large, regular, datasets often dominated by **noise**
- Building on top of existing techniques/teams/collaborations

Software Engineering

- C++/Python
- A few key common roots
  - CASACore, CASA
  - ROOT
- Focus on time-to-solution
- High software complexity
- Complex software engineering
  - Building, version control, collaboration, documentation

High degree of overlap between work-packages

Challenge - example 1

- Takes about 200 days to read one PetaByte **once** if reading from one drive at a time
  - Distribution/parallelisation essential -> drives complexity
- Takes thousands of likelihood evaluations for an accurate Bayesian evidence calculation
  - Each likelihood evaluation typically goes through all the observed data!
- The physical insight comes from connecting multiple datasets, telescopes

Challenge – example 2

- CASA :
  - About 1 million lines of code
  - C++, Fortran, Python, X86 assembly, XML & XSLT, CMAKE, YACC, SWIG, Shell, Perl (probably I forgot some)
  - 35 Direct package dependencies
  - 28 years old already & only recently widespread acceptance

Status

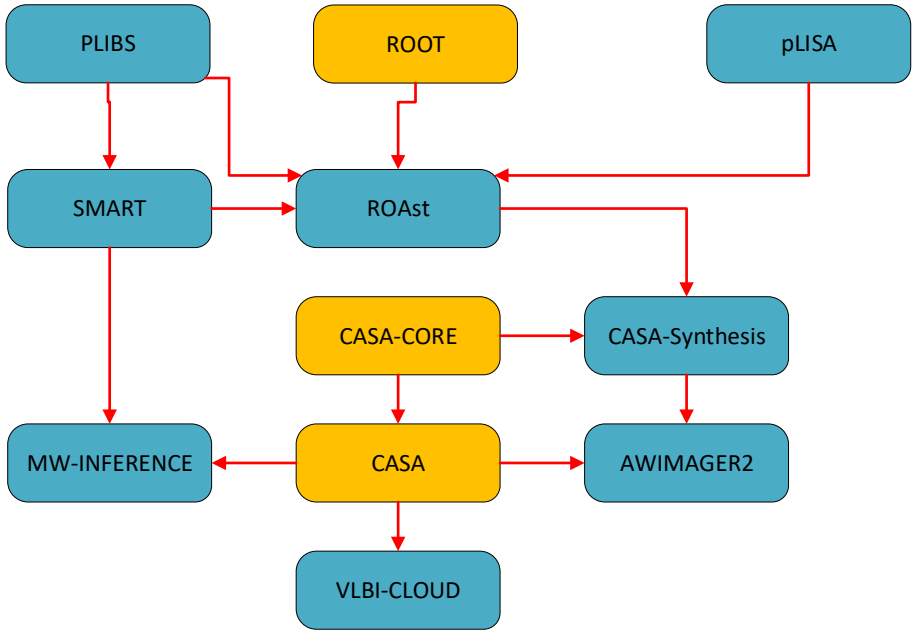
- Work areas, objectives identified ✓
- Initial release of baseline software ✓
- Technology benchmarking report ✓
- Highlights presented here today!

Regular & frequent F2F meetings always helpful !

WIKI – Directory to all publicly available material

- https://www.asterics2020.eu/dokuwiki/doku.php?id=open:wp3:d-ana_software_repository

Potential Links



No "Uses" relationship identified

- DPPP
- CORELib
- STATPLANNER

DATE
28/10/2016

TITLE
D-ANA: Potential "Uses" relationship between modules being developed in the task

DESCRIPTION
View shows how modules being developed within the D-ANA task could use each other. Most of these "use" relationships are not implemented currently (but some are, e.g., CASA-synthesis to AWIMAGER2). This diagram therefore shows the potential high-level architecture that might be achievable at the end of the D-ANA project

Symbols Key

X → Y : Y "uses" X

Module being developed in D-ANA

Module developed/maintained outside ASTERICS

Text for acknowledgement Slide

Acknowledgement

- H2020-Astronomy ESFRI and Research Infrastructure Cluster (Grant Agreement number: 653477).